基于语义关系约束和词语关系信息的句向量研究

夏小强，邵　堃

(合肥工业大学 计算机与信息学院,合肥 230009)

摘　要：针对现有的句向量学习方法不能很好的学习关系知识信息、表示复杂的语义关系，提出了基于 PV-DM 模型和关系信息模型的关系信息句向量模型（RISV），该模型是将 PV-DM 模型作为句向量训练基本模型，然后为其添加关系信息知识约束条件，使改进后模型能够学习到文本中词语之间的关系，并将关系约束模型（RCM）模型作为预训练模型，使其进一步整合语义关系约束信息，最后在文档分类和短文本语义相似度两个任务中验证了 RISV 模型的有效性。实验结果表明，采用 RISV 模型学习的句向量能够更好地表示文本。

关键词：句向量；RISV 模型；PV-DM 模型；关系信息；预训练

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.01.0029

Sentence vector based on semantic relationship constraints and word relationship information

Xia Xiaoqiang, Shao Kun

(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: In view of the fact that the existing sentence vector learning method can not well learn the relational knowledge information and express the complicated semantic relation, this paper proposed a relational information sentence vector model (RISV) based on the PV-DM model and the relational information model. This model used the PV-DM model as the basic model of sentence vector training, and then added the knowledge constraint of relational information to make the improved model can learn the relationship between the words in the text and uses the RCM model as Pre-training model to further integrate the information of the semantic relationship constraints, and finally validates the validity of the RISV model in two tasks: document classification and short text semantic similarity. The experimental results show that sentence vectors learned by RISV model can better represent the text.

Key words: sentence vector; RISV Model; PV-DM Model; Relationship information; Pre-training

0　引言

词向量是一种将词表示为连续向量的技术，是自然语言处理中的一个重要研究课题。自 2013 年 Mikolov 等人[1]提出了 word2vec 模型，词向量在 POS Tagging[2]、句法依存分析[3]、机器翻译[4]以及情感分析等领域取得了丰硕的成果。在大部分的任务中，学习词向量只是工作的第一步，比如说情感分析，需要用学习到的词向量有效的表示文档，这部分是研究工作的另一个难点。

国内外学者在基于词向量的基础上作出了许多重大贡献。唐明等人[5]在 word2vec 模型的基础上，结合 TF-IDF 算法用来表示文档向量。何天翔等人[6]利用大量语料库以及同义词集合构建情感词网，对短文本特征稀疏、信息量不足等问题，提出了结合情感词网的中文短文本情感倾向分析。苗祥等人[7]将特征代表词的同义特征词所对应的情感词加入到该特征代表词的情感词集中,有效提高了特征代表词的情感分析的准确性。Q Liu 等在 15 年提出 SWE 模型[8]，该模型在 skip-gram 基础上组合表示成不等式约束的词语间的关系（同义，上下位等）。Xu 等人[9]则基于 Skip-gram 模型，提出融合关系知识和分类知识的训练框架 RC-NET。2016 年 Nguyen 等人[10]尝试在 Skip-gram 模型基础上加入词汇对比信息共同训练，提出了 DLCE 模型，使得训练得到的词向量能有效识别同义词和反义词。除了基于词向量的工作，Le 等人[11]在 word2vec 模型的基础上添加了 paragraph id，提出了 doc2vec 模型，通过该模型可以直接有效的学习文本句向量。

句向量常见学习方法有求单一文本词向量的平均值[12]，利用 TF-IDF 算法加权后求平均值，对词向量进行聚类[13]，以及使用 doc2vec 模型等。这些方法在一般的学习任务中可以得到不错的结果，但却没有考虑到文本之间语义信息关系和词汇信息，近些年，Liu 等人先后提出了 SWE 模型[8]以及 DLCE 模型

收稿日期：2018-01-17；**修回日期：**2018-03-08

作者简介：夏小强（1990-），男，安徽合肥人，硕士研究生，主要研究方向为自然语言处理、机器学习（yiyele@126.com）；邵堃（1967-），男，副教授，硕士，主要研究方向为开放网络环境下的信任评估模型、需求工程、软件理论，机器学习.

[10]，这些模型尝试在 word2vec 模型基础上，以不同的方式融合不同的结构化信息，取得一定的效果。如 DLCE 模型在同义反义识别任务上表现优异，但在词向量的语义相似性和语义相关性评估任务中，在不同数据集上表现差异较大（SIMLEX999，MEN3000，WS353），模型稳定性不足，SWE 模型在词向量的语义相似性和语义相关性评估任务上有提升，但在同义反义识别任务上却表现不佳。本文在此基础上，借鉴了 SWE 模型添加词语关系信息和 RCM 模型[14]中关系约束的思想提出了本文的关系信息句向量模型（RISV）训练模型，与 SWE 模型相比，本文提出的 RISV 模型在 PV-DM 引入了关系信息，并用关系约束（RCM）模型作为预训练模型，所以能够一定程度的表达复杂的语义关系。最后，在文档分类和短文本语义相似度这两个任务中对模型进行了验证。

1　学习句向量

1.1　RWE 模型

知识图谱中的知识，一般表示为三元组(h,r,t)的形式，其中 r 表示 t 关联的多种不同的关系，例如样本(vegetable, _hyponymy, tomato)。在提取三元组数据后，需要对词语的关系建立表示。例如对于三元组(h,r,t)，若三元组是事实信息，则有 h+r≈t，即 h+r 对应向量应与 t 更相近，该模型称为关系信息模型。模型的输入层是目标词 t 的对应的三元组集合(h,r,t)，投影层做了恒等投影，输出层是在语料中预测目标词。

在 CBOW 语言模型训练中，加入短语关系等信息，使得学习获得的词向量能够很好地表示丰富的语义关系。在此基础上可以得到关系信息词向量模型（relational information word embedding）。目标函数如下：

$$L_{riwe}=\sum_{i=1}^{C}(\log p(w_i\,|\,w_{i-c}^{i+c})+\gamma\sum_{\gamma\in R_{w_i}}\log p(w_i\,|\,h+r)) \qquad (1)$$

其中：函数前半部分是 CBOW 模型目标函数，后半部分是关系信息模型目标函数，γ 是调权参数，C 是训练语料库的大小。

$p(w_i\,|\,h+r)$ 表示已知目标词与词 h 之间存在关系 r，预测目标词为 w_i 的概率，具体计算公式如下：

$$p(w_i\,|\,h+r)=\frac{\exp(e_{h+r}{}^{T}\theta_{w_i})}{\sum_{j=1}^{V}\exp(e_{h+r}{}^{T}\theta_{w_i})} \qquad (2)$$

其中：e_{h+r} 表示向量 e_h 和 e_r 的线性加和，即 $e_{h+r}=e_h+e_r$，θ_{w_i} 表示词 w_i 的分类参数。

1.2　RISV 模型

文本中词语之间具有很多复杂的语义关系，例如上下位关系，在“猫坐在桌子上的垫子里”这个文本中，“猫”是“坐在”的上位词，“桌子”是“坐在”的下位词，这里“坐在”的下位词除了“桌子” 外，还有有“垫子”等，具有相同上位词的“桌子”和“垫子”，从某种意义上来说应该是相似或者相关的，但 Word2vec 模型只是利用大规模语料库中的词语进行训练，所得的词向量只能学习到文本上下文信息，却无法学习到这种词语间的关系，所以其他复杂的语义关系也很难充分表达。

关系信息词向量模型（RWE）主要是基于 CBOW 模型来学习词向量，对于一些任务来说，仍需要将训练好的词向量转换为句向量，所以本文中将关系信息模型引入到 PV-DM 模型，得到关系信息句向量模型（relational Information sentence vector,RISV），RISV 模型目标函数如下：

$$L_{risv}=\sum_{i=1}^{C}(\sum_{d\in D}\log p(w_i\,|\,Cont(w_{i-c}^{i+c},d))+\gamma\sum_{\gamma\in R_{w_i}}\log p(w_i\,|\,h+r)) \qquad (3)$$

其中：d 表示 paragraph id 向量，D 表示 paragraph id 向量空间。

$\mathrm{cont}(w_{i-c}^{i+c},d)$ 表示词 wi 的上下文以及 paragraph id 向量。

使用 Negative Sampling 对 RISV 目标函数进行优化，则对于样本（u,wi）来说，如果 u 的目标词满足条件 E 视为正样本，通过负采样的其他词成为负样本。则指示函数为

$$\delta(w_i\,|\,u)=\begin{cases}1, & w_i=E\\ 0, & w_i\neq E\end{cases} \qquad (4)$$

对于 RISV 目标函数的求解可以分为前半部分和后半部分，则前半部分函数为

$$L_{\mathrm{front}}=\sum_{d\in D}\sum_{i=1}^{C}\log p(w_i\,|\,Context(w_{i-c}^{i+c},d)) \qquad (5)$$

使用随机梯度下降算法对其进行求解可得

$$\theta_u=\theta_u+\eta(\delta(w_i\,|\,u)-\sigma(X_w^T\theta_u))X_w \qquad (6)$$

$$V(w)=V(w)+\eta\sum_{j=2}^{l^w}(\delta(w_i\,|\,u)-\sigma(X_w^T\theta_u))\theta_u \qquad (7)$$

其中：X_w 表示词向量的和或者由词向量连接成，$\sigma(x)=\exp\{x\}/(1+\exp\{x\})$，$\eta$ 为调权参数，V(w)为句向量。L_{front} 函数中（4）式中条件 E 为是否为目标词。

后半部分函数为

$$L_{rear}=\sum_{i=1}^{C}\sum_{\gamma\in R_{w_i}}\log p(w_i\,|\,h+r) \qquad (8)$$

同理，采用随机梯度下降算法进行求解可得：

$$\theta_u=\theta_u+\alpha(\delta(w_i\,|\,u)-\sigma(e_{h+r}^T\theta_u))e_{h+r} \qquad (9)$$

$$e_h=e_h+\alpha\sum_{u\in U}(\delta(w_i\,|\,u)-\sigma(e_{h+r}^T\theta_u))\theta_u \qquad (10)$$

$$e_r=e_r+\alpha\sum_{u\in U}(\delta(w_i\,|\,u)-\sigma(e_{h+r}^T\theta_u))\theta_u \qquad (11)$$

其中：U 表示关系信息集，α 表示调权参数，L_{rear} 函数中式（4）中条件 E 为是否满足关系信息集。则 e_{h+r} 更新可以由公式 $e_{h+r}=e_h+e_r$ 完成。

RISV 模型是在 dco2vec 模型训练中，引入了关系信息知识作为监督，共享词向量，从目标函数来看，PV-DM 模型与关系信息模型线性组合，二者都对句向量更新有着一定影响，调权参数 γ 平衡二者关系，最终使用随机梯度下降算法不断优化参数，得到最优解。

与 RWE 模型相比，本文提出的 RISV 模型可以直接训练出句向量，省略了词向量到句向量转换的过程。另外，RISV 模型添加了 paragraph id 信息，被用来记忆当前文本或文章主题中漏掉的信息，因此造成信息损失较小，在情感分析等任务中相对于 RWE 等模型来说更具优势。

1.3 预训练

在深度学习中，模型预训练起着重要作用[15]。例如，在 Yu M[14]的工作中使用了 CBOW 模型等进行预训练，取得了很好的效果。受 Yu M 等工作的启发，本文使用 RCM 模型进行预训练。

主要思想为假设 Rw 为单词 w 在关系集 R 中的唯一表示。目标是最大化关系语料库 N 中全部单词关系的和：

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{w\in R_{w_i}}\log p(w\,|\,w_i) \qquad (12)$$

其中 $p(w\,|\,w_i)=\exp(X_{w_i}^{T}V_w)/\sum_{\bar{w}}\exp(X_{w_i}^{T}V_{\bar{w}})$，X，V 分别表示输入与输出的词向量。这个模型被称为 RCM 模型。通过 RCM 模型的预训练，然后将预训练词向量作为 RISV 模型的输入，某些参数作为 RISV 模型的初始值。在本文中，RCM 模型可以理解为一种特殊的先验分布带来的正则化，有别于 L1 与 L2 正则化，这种正则化项和 semi-superviesed 以及 early stopping 的原理比较类似。最终的实验效果是给 RISV 模型关系信息的一种补充，为模型增加了关系约束知识。

2 实验与结果分析

实验数据文本语料来自维基百科，爬取数据后，对数据进行去除超链接和中间数据，将数字用数字单词代替等预处理。预处理后总共有一亿个左右单词，然后筛选出出现超过五次的单词，组成包含 202363 个单词的语料库。训练 RCM 模型使用‘词汇’版本的 PPDB（没有短语）语料库，然后筛选出在文本语料库出现的关系对，然后删除重复的关系对，例如，如果<X,Y>包含在 PPDB 中，则删除<Y,X>关系对。三元组语料来自 Freebase，用于关系信息模型的训练。具体信息如表 1 所示。

表 1 实验数据表格

数据	来源	词库	测试集
文本语料	维基百科	202363	
PPDB 语料	PPDB	57829	1583
三元组	Freebase	69023	1657

实验数据部分示例如表 2 所示。

表 2 实验数据部分实例

数据	示例
文本语料	Anarchism is a political philosophy that advocates self-governed societies based on voluntary institutions.These are often described as stateless societies, although several authors have defined them more specifically as institutions based on non-hierarchical free associations.
PPDB 语料	<planning,plans>,<monitoring,monitor>, <seemed,suggested>,<pyramidal,pyramid>
三元组	<The Trail Blazer,is-a,TV Episode>, <The Trail Blazers,Country of origin,United States of America>, <Playing Guitar,is-a,Book>, <Playing Hardball,is-a,TV Episode>

2.1 文档分类任务

2.1.1 实验数据

测试语料库来自 Reuters Corpus,Volume I(RCV1)[16]，该语料库有 806791 个手动分类好的新闻，共有三个大的分类目录，包括主题，工业和地区。本文主要针对主题这个目录的分类，该目录包括四个主题，分类为 C，E，G 和 M，其中 C 表示公司或工业类，E 表示经济类，G 表示政府类以及 M 表示市场类。经过简略的处理后详细信息如表 3 所示。

表 3 文档分类测试语料库信息

类别	训练集	测试集
C	6000	1000
E	1000	500
G	3000	1000
M	3000	1000

2.1.2 实验说明

文档分类任务主要是对语料库中的文档进行分类，由表 3 可知，实验数据总共分为四类，分别为 C，E，G 和 M，语料库总体分为两个部分，分别为训练集以及测试集，训练集用于模型训练，测试集测试实验分类结果，另外，在训练中，会在训练集中分出部分作为作为验证集。具体实验过程如图 1 所示。

由图 1 可知，文档分类任务分为学习句向量和文档分类两个阶段，学习句向量阶段核心是 RISV 模型训练，RISV 模型是在 doc2vec 模型训练中，加入关系信息作为监督信息，二者共享词向量，实际训练中，调权参数 γ 为 0.12 时，实验表现最佳。在文档分类任务中，首先，RCV1 语料库中的文档会在第一阶段训练出句向量，然后，学习到的句向量和相应的类别会组成类似<data,label>形式的数据对做为 SVM 分类器的输入，最后在测试集对模型进行验证，验证方法为给定一组数据<x,y>，如果 SVM 分类器预测类别 y′ 与 y 相同判定分类正确，反之，判定失败，完成所有测试后，汇总计算预测准确率。

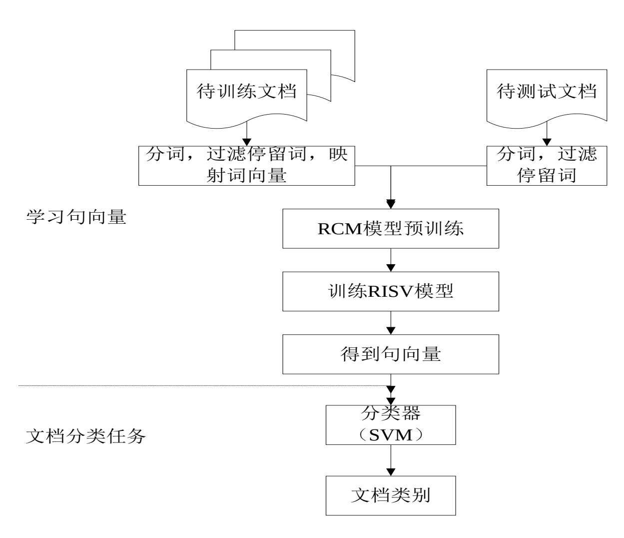


图 1 文档分类任务流程图

2.1.3 实验结果

常见的文档分类方法有基于 word2vec 模型的平均以及 tf_idf 等，基于 doc2vec 模型以及使用 RWE 模型训练词向量进而通过平均以及 tf_idf 等。本文在 RCV1 语料库上进行测试，测试标准为准确率，分类模型为 SVM，验证方式使用五分交叉验证。测试结果如表 4 所示。

表 4 实验结果

模型	准确率（%）				
	C	E	G	M	全部
Word2vec+平均	67.56	69.23	71.25	65.32	68.21
Word2vec+tf_idf	69.26	69.56	70.24	68.67	69.41
Doc2vec	70.35	70.67	69.25	71.26	70.34
RWE+平均	71.25	72.34	71.21	72.39	71.72
RWE+tf_idf	70.35	72.53	**72.36**	72.25	71.77
RISV	72.12	73.21	72.21	72.94	72.54
RISV＋ 预训练	**72.26**	**73.29**	72.34	**72.97**	**72.63**

从表 3 中可以看出，本文提出的 RISV 模型在 C，E 和 M 类别中分类效果比 word2vec 以及 RWE 等方法效果要好。在所有类别汇总的分类中效果也也很明显。另外，在 RISV 模型中使用预训练，从实验数据中可以看出，有一定的提升效果。

2.2 短文本语义相似度任务

2.2.1 实验数据

短文本语义相似度任务使用微软语料库[17]，该语料库常被用来做短文本语义相似度的验证[18~20]。总共有 5 801 个短文本对，每个短文对都用二进制形式来判断语义是否相等。语义相等的短文本对有 3 900 个，不相等的短文本对有 1 901 个，用其中 4 076 个进行训练，1 725 个进行测试。

2.2.2 实验说明

短文本语义相似度任务是计算两个文本的语义相似度。例如，“只有英特尔公司的股息收益率较低”和“只有英特尔 0.3%的收益率较低”语义是相似的，“去年 12 月，他预计增长 5.3%至近 1,540 亿美元” 和 “去年 12 月，他预测增长率为 5%” 语义是不相似的。实验数据使用微软语料库，其数据形式为给定两个短文本，并给出两个文本语义是否相似的判定结果，结果使用二进制表示，1 表示两个短文本语义相似，0 表示两个短文本语义不相似。利用微软语料库，短文本语义相似度任务转换为了二分类问题，具体实验过程如图 2 所示。

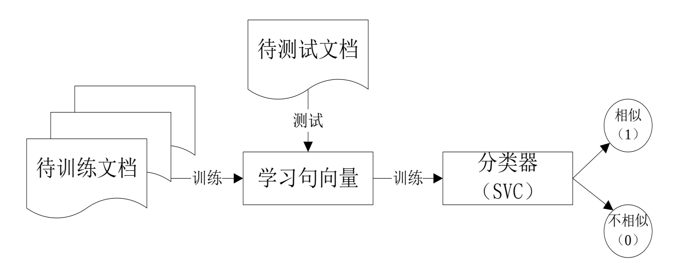


图 2 短文本语义相似度任务示意

由图 2 可知，在短文语义相似度任务中，首先，微软语料库中的文档会被表示为句向量，然后，学习到的句向量会被添加标签 0 和 1，1 表示语义相似，0 表示语义不相似。最后使用测试集测试实验结果，测试方法为给定一组数据<string,label>，如果 SVC 分类器预测相似度 p_label 与 label 相同判定相似，反之，判定不相似，完成所有测试后，计算预测相似准确率。

2.2.3 实验结果

实验使用的评估标准是准确率以及 p(precision)，r(recall) 以及 F_1 值，分类器使用 RBF 核的 SVC 模型，因为特征空间不一定是线性的，验证方式使用五分交叉验证。

表 5 短文本相似度实验结果

模型	准确率	*p*	*r*	F_1
Word2vec+平均	0.6991	0.7123	0.8425	0.7719

Word2vec+tf_idf	0.7012	0.7621	0.8521	0.8046
Doc2vec	0.6929	0.7235	0.9137	0.8076
RWE+平均	0.7102	0.7426	0.8969	0.8125
RWE+tf_idf	0.7201	0.7716	0.9123	0.8361
RISV	0.7312	0.7821	0.9237	0.847
RISV+预训练	**0.7319**	**0.7826**	**0.9314**	**0.8505**

从表 5 中可以看出，RISV 模型在准确率以及 p,r,F_1 值上表现比 word2vec 以及 RWE 等方法好。在预训练后，实验表现能够得到进一步的提升。

2.3 总结

本文主要从两个任务验证 RISV 模型学习句向量的有效性，任务分别为文档分类和短文本语义相似度任务。在文档分类任务中，实验结果如表 6 所示，对 6 个模型学习到的句向量使用 SVM 分类器进行分类，包括对 RISV 进行 RCM 预训练处理。实验结果表明 RISV 模型能够在文档分类任务中取得了很好的表现。在短文本语义相似度任务中，实验结果如表 4 所示，RISV 模型也有很好的表现，并且在两个任务中，使用关系约束模型（RCM）预训练，使初始词向量具有一定关系约束信息，并在 RISV 模型中有一定的体现，从而对实验结果起到帮助。

3 结束语

本文在 RWE 模型的基础上提出了 RISV 模型，与 RWE 模型相比，RISV 模型添加了文档向量，能够记忆段落信息，减少传统从词向量到句向量转换损失的信息，在一个文档训练过程中，paragraph id 保持不变，相当于在预测单词概率时，使用了整个文本的语义。另外，RISV 模型能够直接学习得到句向量，不需要转换。但相对于 RWE 模型以及一些基于 word2vec 模型的算法相比，本文提出的 RISV 模型增加了算法复杂度，在训练中对数据处理也比较繁琐。因此，下一步需要对模型进行改进，优化算法复杂度。

参考文献：

[1] Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc. , 2013: 3111-3119.

[2] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2013: 647-657.

[3] Zhang M, Zhang Y, Che W, *et al.* Chinese Parsing Exploiting Characters [C]// ACL. 2013: 125-134.

[4] Devlin J, Zbib R, Huang Z, *et al.* Fast and robust neural network joint models for statistical machine translation [C]// Proc of Meeting of the Association for Computational Linguistics. 2014: 1370-1380.

[5] 唐明，朱磊，邹显春. 基于 Word2Vec 的一种文档向量表示 [J]. 计算机科学, 2016, 43 (6): 214-217.

[6] 何天翔，张晖，李波，等. 结合情感词网的中文短文本情感分类 [J]. 计算机应用研究, 2015, 32 (10): 2905-2909.

[7] 苗祥，刘业政，孙春华. 领域同义特征词的统计规律及其在情感分析上的应用研究 [J]. 计算机应用研究, 2014, 31 (11): 3333-3336.

[8] Liu Quan, Jiang Hui, Wei S, *et al.* Learning semantic word embeddings based on ordinal knowledge constraints [C]// Proc of the 53th Annual Meeting of the Association for Computational Linguistics. 2015: 1501-1511.

[9] Xu C, Bai Y, Bian J, *et al.* RC-NET: a general framework for incorporating knowledge into word representations [C]// Proc of ACM International Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2014: 1219-1228.

[10] Nguyen K A, Walde S S I, Vu N T. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 454-459.

[11] Le Q V, Mikolov T. Distributed representations of sentences and documents [C]// Proc of the 31st International Conference on Machine Learning. 2014.

[12] Xing C, Wang D, Zhang X, *et al.* Document classification with distributions of word vectors [C]// Proc of Asia-Pacific Conference on Signal and Information Processing Association. 2014: 1-5.

[13] Han K K, Kim H, Cho S. Bag-of-concepts: comprehending document representation through clustering words in distributed representation [EB/OL]. http://dm. snu. sc. kr//static//docs//TR//SNUDM-TR-2015-05. pdf.

[14] Yu M, Dredze M. Improving Lexical Embeddings with Semantic Knowledge [C]// Proc of Meeting of the Association for Computational Linguistics. 2014: 545-550.

[15] Erhan D, Bengio Y, Courville A, *et al.* Why does unsupervised pre-training help deep learning? [J]. Journal of Machine Learning Research, 2010, 11 (3): 625-660.

[16] Lewis D D, Yang Y, Rose T G, *et al.* RCV1: a new benchmark collection for text categorization research [J]. Journal of Machine Learning Research, 2004, 5 (2): 361-397.

[17] Quirk C, Brockett C, Dolan W B. Monolingual machine translation for paraphrase generation [C]// Proc of Conference on Empirical Methods in Natural Language Processing. 2004: 142-149.

[18] Annesi P, Croce D, Basili R. Semantic compositionality in tree kernels [C]// Proc of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014: 1029-1038.

[19] Fernando S, Stevenson M. A semantic similarity approach to paraphrase detection [C]// Proc of Annual Research Colloquium on Computational Linguistics. 2008.

[20] Hu B, Lu Z, Li H, *et al.* Convolutional neural network architectures for matching natural language sentences [C]// Proc of International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2042-2050.